Title

Prediction of Coronary Heart Disease Using Metabolite-based Machine Learning Models

Author
Zhou, Xintong

Publication Date
2021

UNIVERSITY OF CALIFORNIA SAN DIEGO

Prediction of Coronary Heart Disease Using Metabolite-based Machine Learning Models

A thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Electrical Engineering (Machine Learning & Data Science)

by

Xintong Zhou

Committee in charge:

Professor Ramesh Rao, Chair
Professor Mohit Jain, Co-Chair
Professor Farinaz Koushanfar

2021

The thesis of Xintong Zhou is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

For

my mother

my maternal grandmother

my maternal grandfather

my father

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT OF THE THESIS

Prediction of Coronary Heart Disease Using Metabolite-based Machine Learning Models

by

Xintong Zhou

Master of Science in Electrical Engineering (Machine Learning & Data Science)
University of California San Diego, 2021

Professor Ramesh Rao, Chair
Professor Mohit Jain, Co-Chair

Coronary heart disease (CHD) is a leading cause of death in the United States. Currently, the main method of risk assessment is carried out through established risk score algorithms by using traditional risk factors. These algorithms mainly focus on long-term prediction, with the limitation on assessing risk for younger adults. In recent years, with the advancement of serum nuclear magnetic resonance (NMR), more studies of using metabolites to predict CHD have merged. Assessing the risk with metabolites provides

insights into the underlying molecular mechanisms of CHD. This thesis explores that possibility of using metabolites as the predictors and is aiming to understand how much prediction power that machine learning methods could bring in this prediction task.

1. Introduction

Cardiovascular disease is the leading cause of noncommunicable death worldwide, according to WHO, causing 17.5 million deaths in 2012 (World Health Organization, 2014). While cardiovascular disease includes a group of different types of diseases in heart and blood vessels, coronary heart disease (CHD) is the most common type, which is a major cause of death in developed countries (Sanchis-Gomar et al., 2016) and is projected to be the leading cause of death in developing countries too (Okrainec et al., 2004). CHD is a disease that the buildup of plaque in the heart's arteries narrows the coronary arteries over time and then limits or even blocks the blood flood to heart muscles (American Heart Association, 2013). Health professionals may interchangeably use the term coronary artery disease (CAD) and the term coronary heart disease (CHD), which is the result of the former (Sanchis-Gomar et al., 2016).

According to Hajar (2017), "there is still no cure for any form of heart disease". The assessment of the risk becomes more important in prevention practice and in reducing the burden of CHD. Currently, in clinical practice, risk factors are the main metrics being used to predict the risk of CHD. The major risk factors used in the assessment are age, sex, blood pressure, total cholesterol, high-density-lipoprotein cholesterol, low-density-lipoprotein cholesterol, smoking status and diabetes (Wilson et al., 1998). The mainly used algorithms to calculate the risk score are Framingham Risk Score (FRS), which is the most widely used for clinical guidelines, the Systematic COronary Risk Evaluation (SCORE) and the Prospective Cardiovascular Munster (PROCAM) mode (Lloyd-Jones, 2010). These models are for the prediction of the risk of CHD in 10 years, which is a "substantial improvement over clinical judgement alone for appropriate risk stratification" ((Lloyd-Jones, 2010). However, there is certain limitation of these 10-year risk prediction models. Because these models are heavily dependent on age, for younger adults (men <45 years of age and women <65 years of age, moderate elevation in risk factors would only reflect little effect in the 10-year risk prediction (Cavanaugh-Hussey et al., 2008). This means that even for younger adult has a substantial life-time risk in CHD, the 10-year risk assessment would still categorize him or her into the low-risk interval.

Besides the risk factors, multiple studies have indicated that CHD risk assessment could be improved by utilizing novel biomarkers, such as coronary artery calcium and cardiac troponim-T, but only a few tractable biological pathways have been identified (Wang et al., 2019). However, the recent development of nuclear magnetic resonance (NMR) has enabled faster and low-cost metabolite detection (Miggiels et al., 2019). With this innovation, more studies have been conducted to understand the possibility of predicting CHD by using metabolites. Metabolites are small molecules produced during metabolism and the end-product of gene expression (Forssen et al., 2017). There are studies of understanding the association between metabolomic profile and CHD (Vaarhorst et al., 2014) (Wang et al., 2019). These studies demonstrate that metabolites are a promising tool to improve the prediction of CHD. The feasibility of predicting CHD by using metabolite can present a molecular level reflection, which has the potential to enable professionals to understand CHD by the metabolic pathway. Additionally, due to the cause of CHD, metabolites may provide insights into the chronic changes of the artery walls. In this research, machine learning algorithms are utilized to predict the development of CHD within five, 10 and 15 years.

2. Related Work

Currently, there are multiple studies in applying machine learning methods to predict CHD using risk factors and in investigating the association between metabolites and CHD by survival analysis. However, there are not many studies or research on utilizing machine learning to predict CHD with metabolites in searching related work.

Forssen et al. (2017) investigate the prediction performance for CHD by using three machine learning algorithms, logistics regression, principal components analysis (PCA) and random forest. They use Clinical Cohorts in Coronary disease Collaboration (4C) as the dataset, which are collected through UK NHS hospitals. In their studies, the occurrence of CHD is defined if more than 50% stenosis can be found in more than one coronary arteries, which is different from the interest of this research, a timely-basis prediction. Their studies can help to associate metabolites with CHD but not in a timely manner. They use PCA as a feature selection method and the two algorithms, logistic regression and random forest, as the classifier. They first apply PCA on 256 metabolites and select the first six components that account for more than 95% data variability as the selected metabolite predictors. Then they implement the selected six PCA-derived metabolite factors into logistic regression with L1 penalization and random forest. For comparison, they also run these two models including both the six PCA-derived metabolite factors and four risk factors (age, sex, use of statins, hypertension) as their adjusted models. In the result, random forest has both a better

AUC (0.675) and accuracy (0.713) than the logistic regression with PCA-derived features whose AUC is 0.625 and accuracy is 0.686. However, for the adjusted models, the adjusted logistic regression (0.767 AUC and 0.759 accuracy) outperforms the adjusted random forest (0.711 AUC and 0.732 accuracy). Using PCA as feature selection is a good way to improve model performance, however, losing the interpretability, since PCA is to project features into lower dimension. Their studies show that risk factors can improve the prediction of using metabolite only.

In the work of Vaarhorst et al. (2014), prediction of CHD using metabolites is assessed and compared with using traditional risk factors (lipid levels, blood pressure, lifestyle factors, family history, sex and age). They use logistic regression with LASSO to select relevant metabolites associated CHD and compare with the prediction performance of using metabolites that are not associated with risk factors. They find out that using metabolites that are independent of traditional risk factors could not improve risk prediction based on traditional risks in the groups of people who are free from CHD. They infer that the performance is due to the reason that many of metabolites in their data are dependent of risk factors. Those dependent metabolites are demonstrated in some studies that have important weights in predicting CHD. Since they compare the risk prediction of using only metabolites independent of risk factors, the lack in the quantities of metabolites is the main reason that the performance could not be beyond that of using risk factors.

Yu et al. (2020) take a retrospective study on finding the significant risk factors of metabolic syndrome by utilizing machine learning algorithms. In their research, they consider every known risk factor of metabolic syndrome and utilize machine learning algorithms to select the significant ones based on the prediction performance. Metabolic syndrome is a cluster of disorders that affect the development of numerous diseases, including cardiovascular disease. They sample the data from all the Taiwanese aged 18 and above who utilize FibroScan for self-health examination. The data is collected from the score generated by FibroScan, a non-invasive device to assess the hardness of liver and generate relevant metric scores. In their study, they use the definition of metabolic syndrome by National Cholesterol Education Program Adult Treatment Panel II to identify if a participant has the metabolic syndrome or not. In total, there are 193 individuals diagnosed with metabolic syndrome and 1140 without. They use multiple algorithms for this study, including classification tree, Chi-square, random forest, generalized linear model and logistic regression. Their results show that random forest has the best accuracy because the model removes the bias that a decision tree model might have and thus improving the predictive power. They discuss that one limitation of their study is that as a retrospective one, it is not sufficiently powerful to demonstrate the usefulness of machine learning in diagnose metabolic syndrome if under a prospective study. Another limitation is that their data population is all from people who are using a self-check device for possibility of metabolic syndrome, which could be bias and should be validated in other populations.

3. Dataset

The dataset used in this research is derived from FINRISK, a population-based study for noncommunicable disease monitoring and intervention of Finland, coordinated by the National Institute for Health and Welfare in Finland from 1972 to 2012. Data in FINRISK are collected from the questionnaires, health examination and blood samples of each participants at baseline, who are followed up every 5 years and collected all these three types of in each follow-up year (Borodulin et al., 2017). In 1992, high throughput profiling of circulating metabolites is collected from participants via NMR (Delles et al., 2017). With a systematic and standardized data collection procedure for 40 years, the data of FINRISK provides professionals insights into understanding noncommunicable diseases and approaches to promote public health.

The participants in FINRISK are from age of 25 to 74 years old in selected areas of Finland, including North Karelia, Northern Savo, Turku and Loimaa, Helsinki and Vantaa, Northern Pohjanmaa and Kainuu and Lapland (Borodulin et al., 2017). For the purpose of the research work in this thesis, 7643 participants (N=7643) are used with10 risk factor features (Table 3.1) and 247 metabolite features as predictors, all of which are collected at baseline from FINRISK. The class labels in this research are derived by using two features, 'CHD' and 'CHD_time'. 'CHD_time' is a non-negative continuous value, presenting the final follow-up time. If CHD occurs, 'CHD_time' is recorded and stopped. 'CHD' is a discrete value, indicating if CHD occurs or not at 'CHD_time': 1 meaning that CHD occurs and 0 meaning not. The class label is defined by if there is a CHD event within given years. Table 3.2 is an example of defining the class label within 5 years (CHD_time ≤ 5). In this research, for comparison, we consider the case if CHD occurs within 10 years, a time length that is used the current risk score algorithm. For exploration of the prediction power of the machine learning models, cases of 'CHD_time' is five and 15 years are also tested. Table 3.3 shows the count of CHD event of each class within five, 10 and 15 years.

Table 3.1 Risk factor features used in the research

Name	**Type**	**Description**
Age	Continuous	Ranging from 25 to 74 years old
Sex	Discrete	0 = Female, 1 = Male
BMI	Continuous	Body Mass Index, ranging from 15.8 to 53.5
Diabetes	Discrete	Diabetes status, 0 = No, 1 = Yes
Smoking	Discrete	Smoking status, 0 = No, 1 = Yes
LDL	Continuous	Low-density lipoproteins (mmol/L)
HDL	Continuous	High-density lipoproteins (mmol/L)
TG	Continuous	Triglycerides (mmol/L)
Hypertension	Discrete	Hypertension status, 0 = No, 1 = Yes
sysBP	Continuous	Systolic blood pressure (mmHg)

Table 3.2 Examples of defining class label for the occurrence of CHD in 5 years

CHD_time (years)	**CHD** (0 = No, 1 = Yes)	**Class Label** (1: CHD = 1 & CHD_time<=5)
4.7	1	1
6	1	0
6	0	0

Table 3.3 Counts of CHD event within five, 10, and 15 years

CHD	**Within 5 years**	**Within 10 years**	**Within 15 years**
Not occur (0)	7538	7395	7236
Occurs (1)	105	248	407

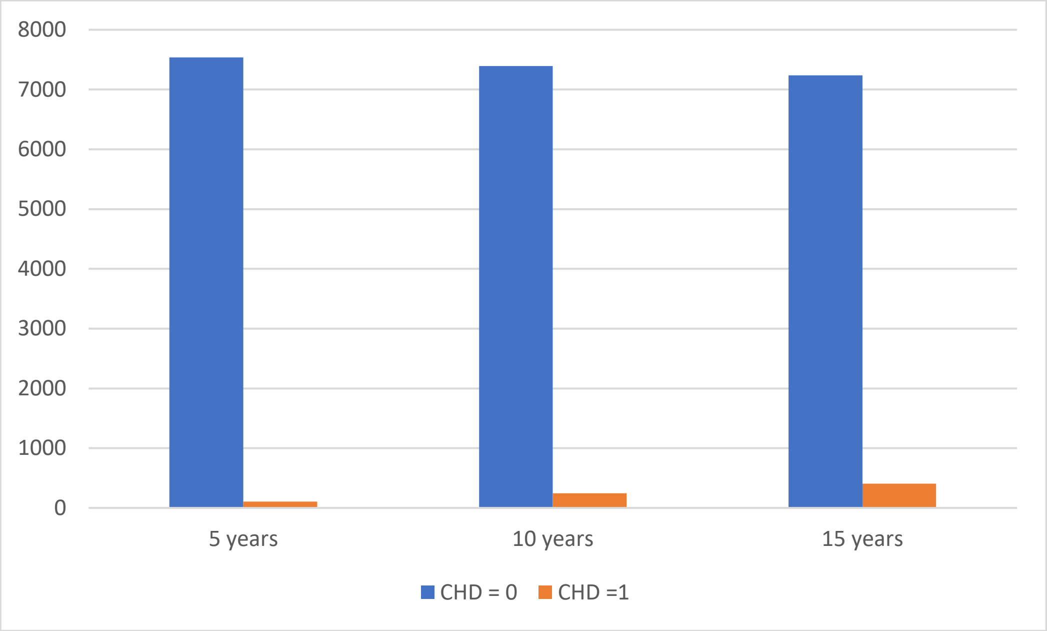

Figure 3.1 Distribution of CHD event within 5, 10 and 15 years

4. Method

4.1 Data Preprocessing

Data preprocessing is to ensure that the data is consistent and ready to implement in machine learning models. Three major aspects are checked and prepared for this dataset: missing data, data standardization, and class imbalance.

Missing Data

For missing data, dropping and imputation are used based on features. For the risk factor data, there are 138 out of 7643 participants with missing data in different features: ‘diabetes’, ‘smoking’, ‘LDL’, ‘hypertension’, and ‘sysBP’. These 138 participants are dropped directly. Some of the features are from personal response to questionnaire, such as ‘smoking’ and ‘diabetes’, which is nearly impossible to infer without asking the participant. Some features are obtained from the results of health examination and blood sample, such as ‘hypertension’ and ‘LDL’, which may have some underlying relationship. Imputing these data without the certainty of knowing that relationship may undermine the model performance and gives misguided results. For these reasons above, participants with missing data in the risk factor features above are dropped. For the metabolite data, missing data is replaced by 0 because it is highly possible that the missing metabolite data is due to the low values of metabolite detected.

Data Standardization

Some machine learning models are optimized based on distance, such as K-nearest neighbors. Some are optimized by applying gradient descent, such as support vector machine. Data standardization could enable features in different units to contribute equally to the model, such as ‘age’, ranging from 25 to 74, and ‘BMI’, ranging from 15.8 to 53.5. In this research, data is scaled by applying z-score standardization, which is by removing the mean and scaling to unite variance.

Class *Imbalance*

Class imbalance is when the number of events in each class is not equal. The class with more events is called the majority class. The class with fewer events is called the minority class. Class imbalance becomes an issue if the minority class has too few in the training sample or has too few events compared to the majority class, such as 1 to 100. In this research, the class of not having CHD event is over 100 times more than the class of having CHD, the class of interest. With such imbalanced data, the classification models would have difficulty in distinguishing the minority class. As a result, the classifiers would have poor predictive ability for the minority class and has the tendency to classify most new samples into the majority class. To alleviate the imbalanced class issue, an oversampling method, Synthetic Minority Oversampling TEchnique (SMOTE) is used. SMOTE is an oversampling technique to randomly generate synthetic samples, linear combination of the two similar samples from the minority class. SMOTE is a widely used and efficient method to reduce the class imbalance issue after feature selection (Blagus & Lusa, 2013). Given the sample size, 7643 and the class size ratio, 1 to 100, an oversampling method is more preferrable than an undersampling method, which discards the majority class till the skewness is less severe, such as 1 to 10.

4.2 Feature Selection

Feature selection is a method to select the relevant features to the prediction target and to discard the irrelevant features. It can reduce the irrelevant information and overfitting. Also, with a smaller feature space after feature selection, not only the computation performance can be improved but the possibility of curse of dimensionality could also be avoided. Curse of dimensionality is an issue that in a high-dimensional feature space, the features are so sparse that the classifier has a poor performance to learn about the relationship between features and the classification target. In this research, two feature selection methods are used for metabolite data, ANOVA and LASSO. Both methods are to select features that are relevant to the classification target. After comparing the overall performance of all the models by using two feature selection methods, ANOVA is used for further analysis.

ANOVA

ANOVA, Analysis of Variance, is a technique to determine if a feature could well separate two classes by examining if there is a significant different in the means of two or more samples. It is a filter method in feature selection, which selects the features by ranking the usefulness or correlation of features to the model, based on F-statistics (Fonti & Belitser, 2017). In this research, top 50 features based on ANOVA are selected for each of 5, 10 and 15 years. Then, the common ones among these three sets of 50 features are chosen. There 32 metabolites in common. This is because we would like to know how well one set of metabolites can do to make the prediction.

4.3 Modeling

Four supervised learning models are implemented for this classification task, logistic regression, support vector machine (SVM), random forest and K-Nearest Neighbors (KNN). These four models are the common classifiers that have been investigated in others' research work for predicting heart diseases or related health condition using metabolite data. Forssen et al. (2017) tests the performance of logistic regression in the prediction of CHD using metabolite data. Gutiérrez-Esparza et al. (2020) implements random forest to predict metabolic syndrome, "a health condition that increases the risk of heart disease". Decision tree, a unit in random forest, SVM and KNN are investigated and compared by Pouriyeh et al. (2017). Although their research work uses different data source to implement these models, the modeling results show that these four classifiers have provided useful insights in terms of model performance and interpretation.

Logistic Regression (LR)

Logistic regression is an algorithm that uses sigmoid function to transfer a linear regression model into a binary classification model. With the sigmoid function transformation, the predicted result of a linear regression is scaled to the range from 0 to 1. This scaled result represents the odds of the class of interest. Then, a certain threshold is set to determine which class the model predicts the data into based on the odds.

Support Vector Machine (SVM)

SVM is a very popular supervised learning machine learning algorithm for classification (Ramalingam et al., 2018). SVM finds a hyper-plane in the feature space, a plane that separates classes with an optimization in the margin distance between the data points to the hyper-plane. For example, in the case of two classes, SVM works on finding a hyper-plane that could separate two classes so that the distance between the data point closet to the hyper-plane is the widest for both classes. SVM uses kernel functions to project data into a higher dimensional space where the data is separable (Campbell & Ying, 2011). In this research, radial basis function kernel is used as it could form a non-linear classifier, which is different from the linear classifier formed by logistic regression.

Random Forest (RF)

Random forest is an ensemble classifier consisting of multiple decision trees. Decision tree is a simple tree-based classification algorithm. It has three types of nodes, chance node, decision node, and end node. A chance node shows the possible outcomes of a particular node; a decision node is where a decision is made based on the outcome; an end node is a node that gives the final decision. A decision tree starts from a root node and splits into various nodes and branches based on Gini index and entropy rule. Each node derives some information on the features. Each link indicates a decision rule to the following node (Krishnani et al., 2019). However, decision tree may have overfitting issue due to its mechanism of nodes and decision making. Random forest is a remedy to that overfitting issue. It uses bagging approaching and learns the model based on the overall performance of all the decision trees. The general institution of random forest is that more decision trees would make a better decision. It produces manifold decision trees and decides the class based on a voting system, choosing the class that most decision trees agree on.

K-Nearest Neighbors (KNN)

KNN is a nonparametric technique for pattern classification. It makes no assumption on the data, which is useful when there is no prior knowledge of the data distribution. KNN predicts the class of a data point based on how similar that data point is to the data with class labels in the training dataset. In other words, this algorithm compares the characteristics of a data point without a label to those with class labels in the training set. KNN computes the Euclidean distance from each feature of an unclassified data point to the features of classified data and selects the k closet classified data points (neighbors). Then, the unclassified data will be predicted as the class that has the most counts in these k selected neighbors.

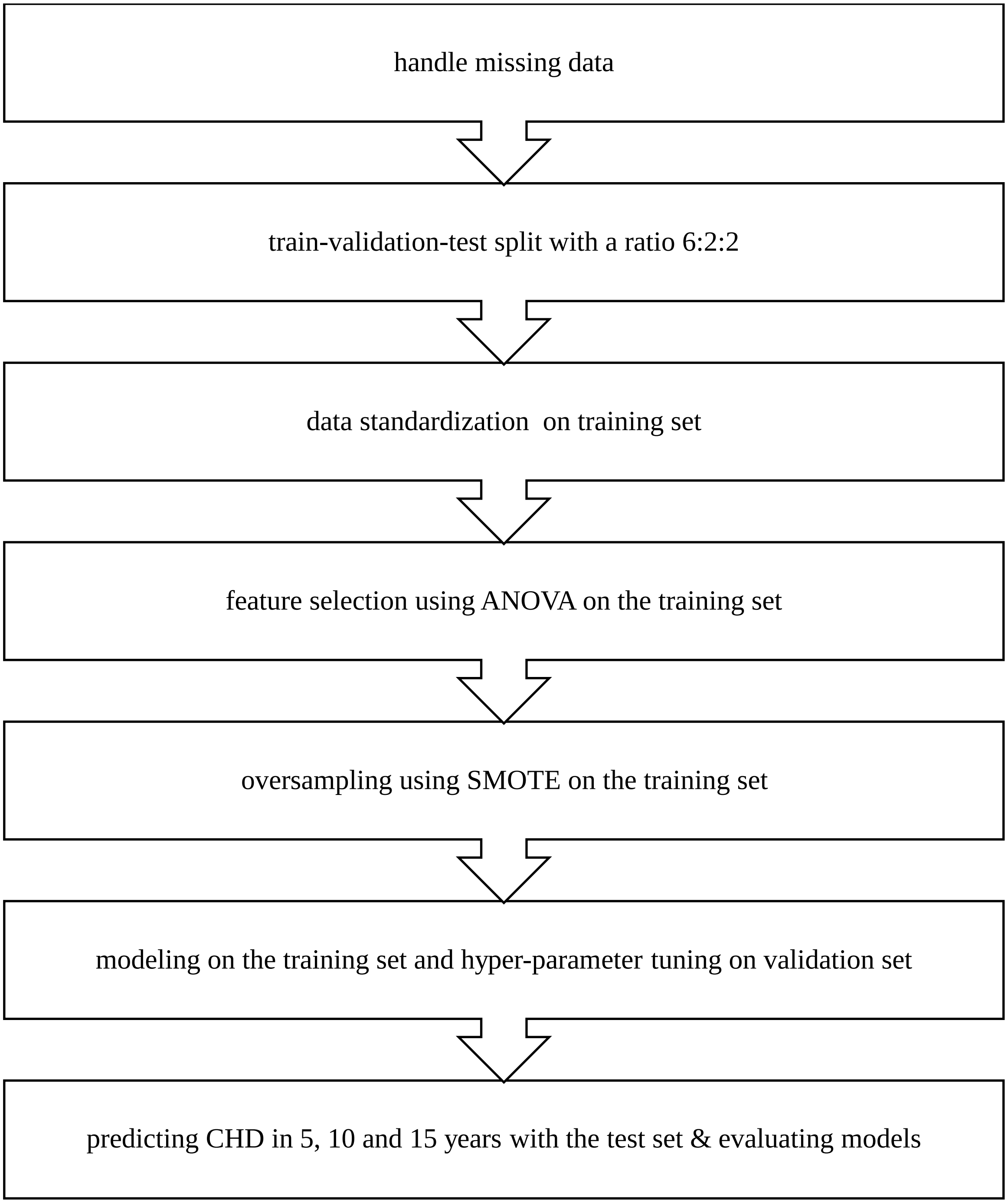
handle missing data
train-validation-test split with a ratio 6:2:2
data standardization on training set
feature selection using ANOVA on the training set
oversampling using SMOTE on the training set
modeling on the training set and hyper-parameter tuning on validation set
predicting CHD in 5, 10 and 15 years with the test set & evaluating models

Figure 4.1 Flowchart of proposed workflow

5. Results

AUC, F2 score and accuracy are the major metrics used to evaluate the model performance. In this research, the dataset suffers from server imbalanced class. The SMOTE oversampling method can alleviate the imbalanced class issue but cannot enable the model to have the same metrics performance as the balanced class (Blagus & Lusa, 2013). Thus, the major metrics chosen for evaluation may have some differences from the common ones, such as accuracy. Several other metrics are also produced in the model evaluation but are not used as the main metrics for evaluation. Below is the definition of each metrics produced.

True Positive (TP): It is an outcome when the positive class is correctly classified as positive by the model

True Negative (TN): It is an outcome when the negative class is correctly classified as negative by the model

False Positive (FP): It is an outcome when the negative class is incorrectly classified as positive by the model

False Negative (FN): It is an outcome when the positive class is incorrectly classified as negative by the model

Accuracy: It is the ratio of number of correct predicts given by the total instances. In data with imbalanced class, accuracy may not be a good indicator for model evaluation due to its bias towards the majority class. For example, in a dataset with 1000 negative class and 100 positive class, a model that predicts 970 negative class and 30 positive class correctly would give a high accuracy of 91% but only a low recall (defined in the following) of 0.3.

However, accuracy is a commonly used evaluation metrics even in imbalanced class modeling because it is the most intuitionistic metrics to interpret in classification (Haixiang et al., 2017).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision: In this research, it measures the proportion of individuals who develop CHD and are correctly predict out of all the individuals who are predicted to develop CHD

$$Precision = \frac{TP}{TP + FP}$$

Recall: In this research, it measures how well the model could correctly predict the individuals who would develop CHD. It is the proportion of individuals who develop CHD and are correctly predicted out of all the individuals who develop CHD.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score: It is the harmonic mean of precision and recall.

$$F1\ Score = 2 * \frac{Precsion \times Recall}{Precision + Recall}$$

F2 Score: It is based on F1 score but gives more weight to recall. In this research work, the class of interest (positive class) suffers from severe imbalanced class issue. Oversampling method can only alleviate the issue but cannot completely enable the model to have the same metrics performance as the balanced class. Thus, F2 score can provide more insights into the positive class for the imbalanced class data.

$$F2\ Score = (1+2^2) * \frac{Precsion \times Recall}{2^2 \times (Precision + Recall)} = 4 \times \frac{Precsion \times Recall}{5 \times (Precision + Recall)}$$

Receiver Operator Characteristics (ROC): is a probability curve illustrating the capability that the model distinguishes two classes. ROC curve is a plot of the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) at different threshold. Area Under the ROC Curve (AUC) is a numerical measure ranging from 0 to 1. The closer the value of AUC to 1, the better the model distinguishes two classes. AUC is one of the most important metrics for model evaluation in this research. Although AUC is questionable by Hand (2009) for its consistency in measuring imbalanced class, it is still a useful and the most used metrics to measure model performance with some counter arguments to Hand's (Haixiang et al., 2017)

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

In the task of predicting the development of CHD, we would like to see if we can find one set of metabolites that can predict CHD in five, 10 and 15 years compared to the performance of using risk factors. As using ANOVA is the main feature selection strategy, there are several different ways of selecting metabolites: selecting top K (K = 15, 20, 25, 30) for prediction model for each of those three years' timeline; selecting the common metabolite features among top N (N=30, 40, 50) from each of three years' timeline; using common top N metabolites plus risk factors; selecting K top features in the combination of metabolites and risk factors. Figure 5.1 to Figure 5.10 and Table 5.1 to 5.9 are examples of model results of using different feature selection strategies. There is no feature selection of risk factors because these 10 risk factors are the common ones used in existing risk assessment system.

In the result of implementing the model with the selected 25 metabolites, logistic regression has the best overall performance in terms of AUC with random forest followed, which has a similar performance with logistic regression. If ranking the performance by F2-score, logistic regression still has the best overall performance but followed by SVM, which has a similar performance with logistic regression. The trend also appears in the result of predicting the development of CHD by using only risk factors and the result of predicting by using the combination of the selected top 25 metabolites and the risk factors. In predicting the development of CHD in 5 years, AUC of logistic regress is a little bit less than that of random forest and a little bit more than that of SVM; F2-score of logistic regression is a little bit less than that of SVM and a little bit more than random forest. For the prediction of the development of CHD in 10 and 15 years, logistic regression has the best performance in terms of both AUC and F2-score. According to the result, it can be observed that random forest has the highest accuracy but the lowest recall, which means that random forest has an outstanding performance in classify the negative class, which is the no occurrence of CHD in this research. Thus, random forest has a lower AUC than logistic regression because AUC is dependent on TPR and FPR, a rate that has negative class as the denominator. The less outstanding performance of SVM could be due to the method of oversampling. Since SVM is an algorithm to find an optimal hyper-plane to segregate two classes, the oversampling method of SMOTE may have effect on the performance of SVM. In all the prediction tasks, KNN has the lowest AUC score but the recall of KNN varies in predicting for different years:

some is the highest, some is the second or third highest. Overall, the ranking of the performance of each model varies by different evaluation metrics. If considering the integrated performance of the metrics, logistic regression and SVM are the two best models in the prediction of CHD in five years by using metabolites; logistic regression is the best model in the prediction of CHD in five, 10 and 15 years by using metabolite or risk factors or the combination of these two sets of features.

Overall, how many features are selected appears to have similar performance within the same type of set of features (metabolites only, risk factors only, metabolites and risk factors). Prediction of combing metabolites & risk factors show close or a little bit better performance than using risk factors, but risk factors have important roles in using metabolites and risk factors together. Prediction with metabolites has the least outstanding performance but the performance is getting closer to risk factors in longer timeline. Among all these models, logistic regression has the best performance. However, the differences are not significantly large for some

Table 5.1 Model performance results with 25 selected metabolites in 5 years

Model	AUC	F2 score	F1 score	Accuracy	Recall	Precision
LR	0.71	0.11	0.05	0.74	0.53	0.03
SVM	0.70	0.15	0.08	0.90	0.37	0.05
RF	0.78	0.10	0.09	0.97	0.11	0.07
KNN	0.62	0.1	0.04	0.69	0.58	0.02

Table 5.2 Model performance results with 25 selected metabolites in 10 years

Model	AUC	F2 score	F1 score	Accuracy	Recall	Precision
LR	0.72	0.19	0.10	0.71	0.53	0.05
SVM	0.65	0.16	0.08	0.61	0.58	0.04
RF	0.69	0.10	0.08	0.93	0.11	0.07
KNN	0.61	0.12	0.06	0.71	0.33	0.04

Table 5.3 Model performance results with 25 selected metabolites in 15 years

Model	AUC	F2 score	F1 score	Accuracy	Recall	Precision
LR	0.76	0.32	0.18	0.68	0.69	0.10
SVM	0.67	0.23	0.14	0.78	0.36	0.09
RF	0.71	0.10	0.09	0.91	0.09	0.09
KNN	0.66	0.25	0.13	0.64	0.64	0.07

Table 5.4 Model performance results with risk factors in 5 years

Model	AUC	F2 score	F1 score	Accuracy	Recall	Precision
LR	0.82	0.16	0.07	0.74	0.84	0.04
SVM	0.69	0.11	0.05	0.86	0.32	0.03
RF	0.76	0	0	0.98	0	0
KNN	0.65	0.11	0.05	0.81	0.42	0.03

Table 5.5 Model performance results with risk factors in 10 years

Model	AUC	F2 score	F1 score	Accuracy	Recall	Precision
LR	0.80	0.27	0.14	0.71	0.78	0.08
SVM	0.73	0.21	0.12	0.78	0.49	0.07
RF	0.75	0.07	0.07	0.95	0.07	0.07
KNN	0.67	0.17	0.09	0.75	0.42	0.05

Table 5.6 Model performance results with risk factors in 15 years

Model	AUC	F2 score	F1 score	Accuracy	Recall	Precision
LR	0.8	0.36	0.20	0.70	0.74	0.12
SVM	0.71	0.3	0.18	0.74	0.56	0.11
RF	0.75	0.14	0.14	0.92	0.13	0.14
KNN	0.67	0.26	0.14	0.70	0.49	0.08

Table 5.7 Model performance with 18 common ones out of 25 selected metabolites & risk factors in 5 years

Model	AUC	F2 score	F1 score	Accuracy	Recall	Precision
LR	0.80	0.14	0.07	0.76	0.68	0.03
SVM	0.70	0.10	0.07	0.95	0.16	0.04
RF	0.78	0	0	0.98	0	0
KNN	0.73	0.10	0.05	0.72	0.58	0.03

Table 5.8 Model performance with 18 common ones out of 25 selected metabolites & risk factors in 10 years

Model	AUC	F2 score	F1 score	Accuracy	Recall	Precision
LR	0.80	0.27	0.14	0.74	0.71	0.08
SVM	0.76	0.21	0.14	0.88	0.33	0.09
RF	0.80	0.03	0.03	0.96	0.02	0.05
KNN	0.70	0.22	0.11	0.66	0.69	0.06

Table 5.9 Model performance with 18 common ones out of 25 selected metabolites & risk factors in 15 years

Model	AUC	F2 score	F1 score	Accuracy	Recall	Precision
LR	0.80	0.36	0.2	0.71	0.74	0.12
SVM	0.71	0.27	0.19	0.82	0.40	0.12
RF	0.75	0.09	0.11	0.94	0.08	0.18
KNN	0.70	0.29	0.16	0.62	0.70	0.09

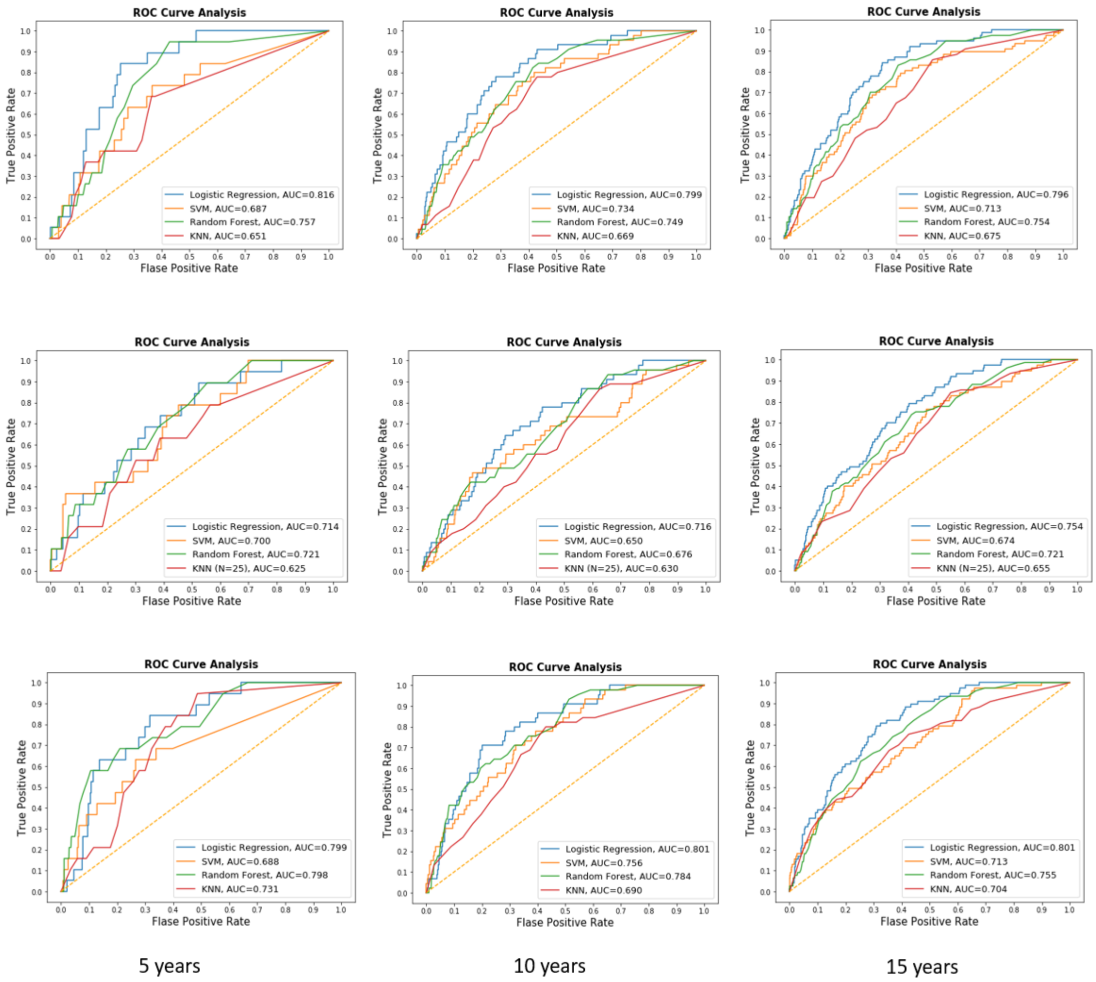

Figure 5.1 ROC curves for 4 machine learning models. Each row is the prediction of CHD in 5, 10 & 15 years from left to right. The top row is prediction with risk factors. The middle row is prediction with top 25 selected metabolites. The last row is prediction with top 25 selected metabolites & risk factors

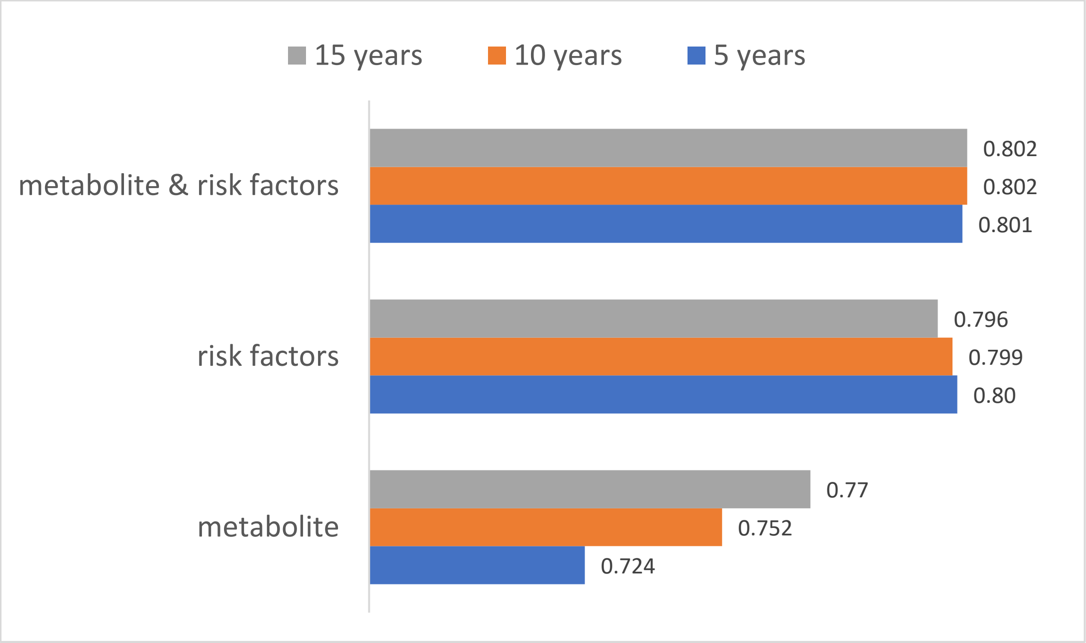

Figure 5.2 Average AUC of the best performed models for 3 different sets of features: metabolite, risk factors, metabolite & risk factors, in 5-, 10- and 15-year prediction of CHD

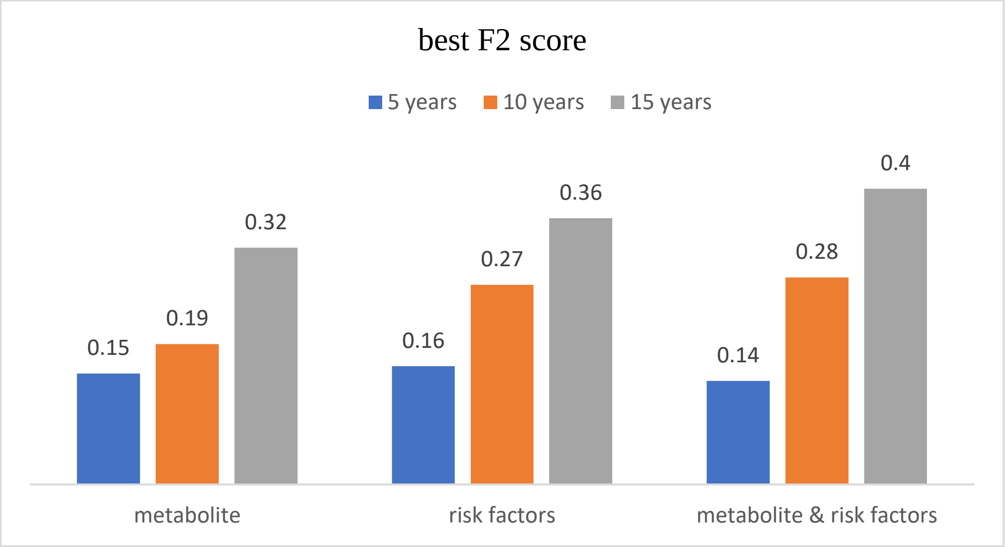

Figure 5.3 Best F2 score for 3 different sets of features: metabolite, risk factors, metabolite & risk factors, in 5-, 10- and 15-year prediction of CHD

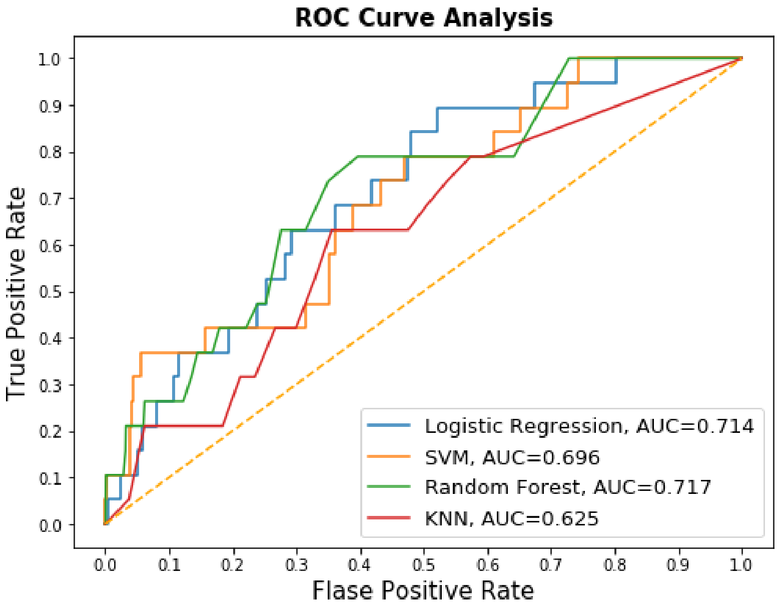

Figure 5.4 ROC curves for 4 machine learning models with top 25 selected metabolites for prediction of CHD in 5 years

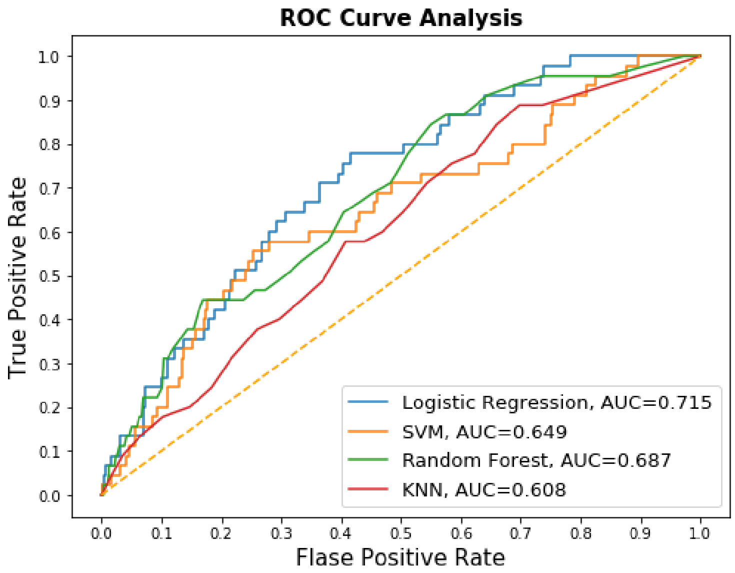

Figure 5.5 ROC curves for 4 machine learning models with top 25 selected metabolites for prediction of CHD in 10 years

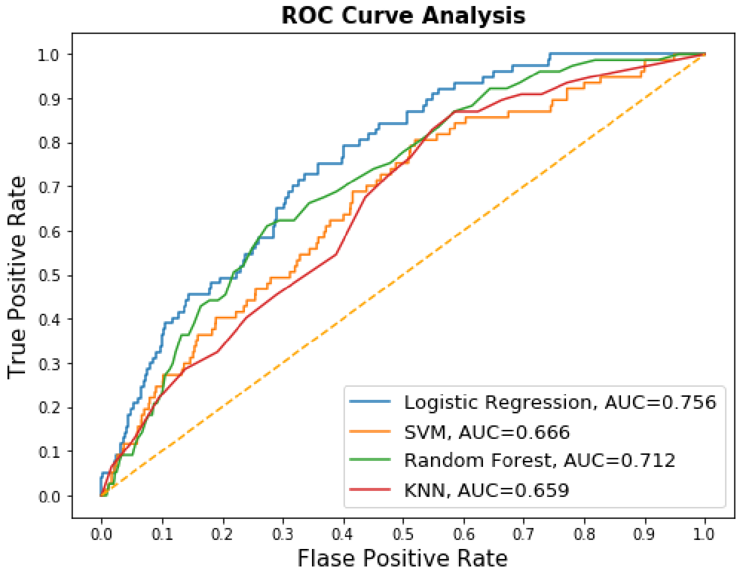

Figure 5.6 ROC curves for 4 machine learning models with top 25 selected metabolites for prediction of CHD in 15 years

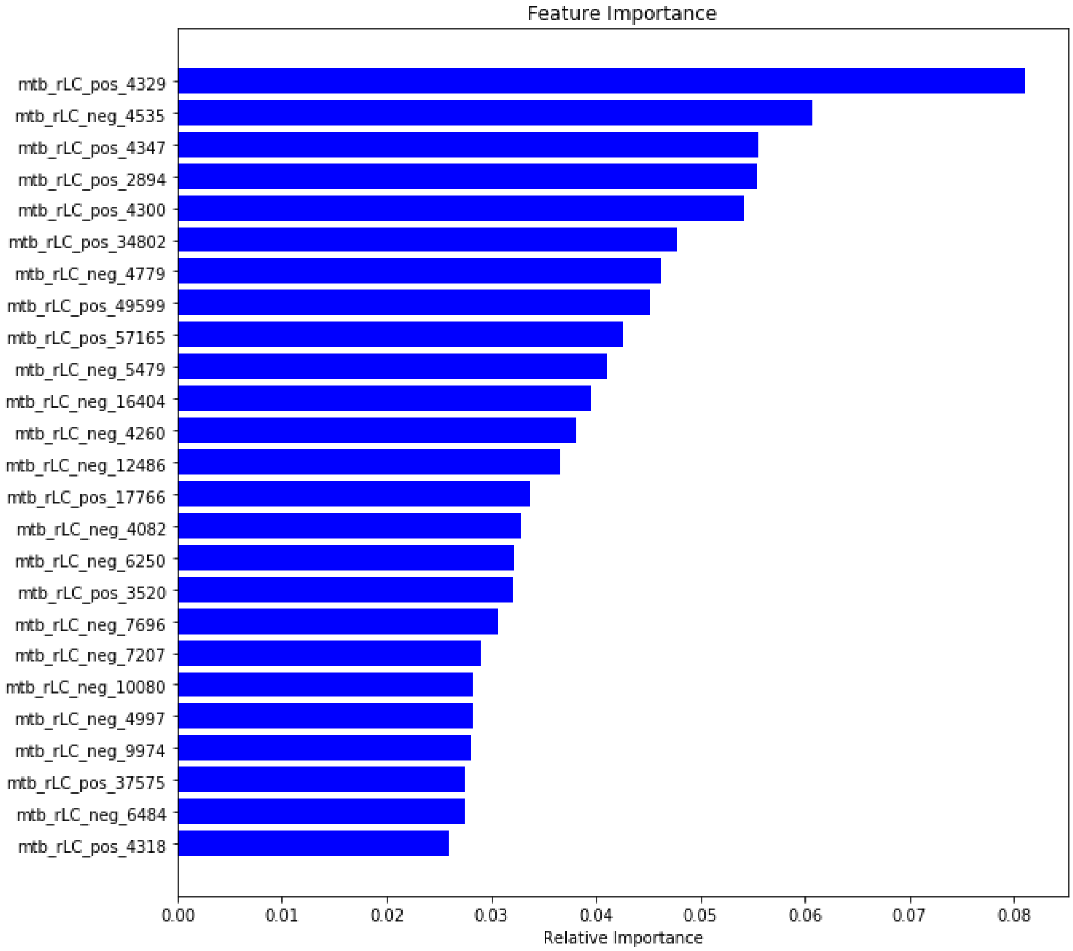

Figure 5.7 Feature importance from random forest of top 25 selected metabolites for prediction of CHD in 5 years

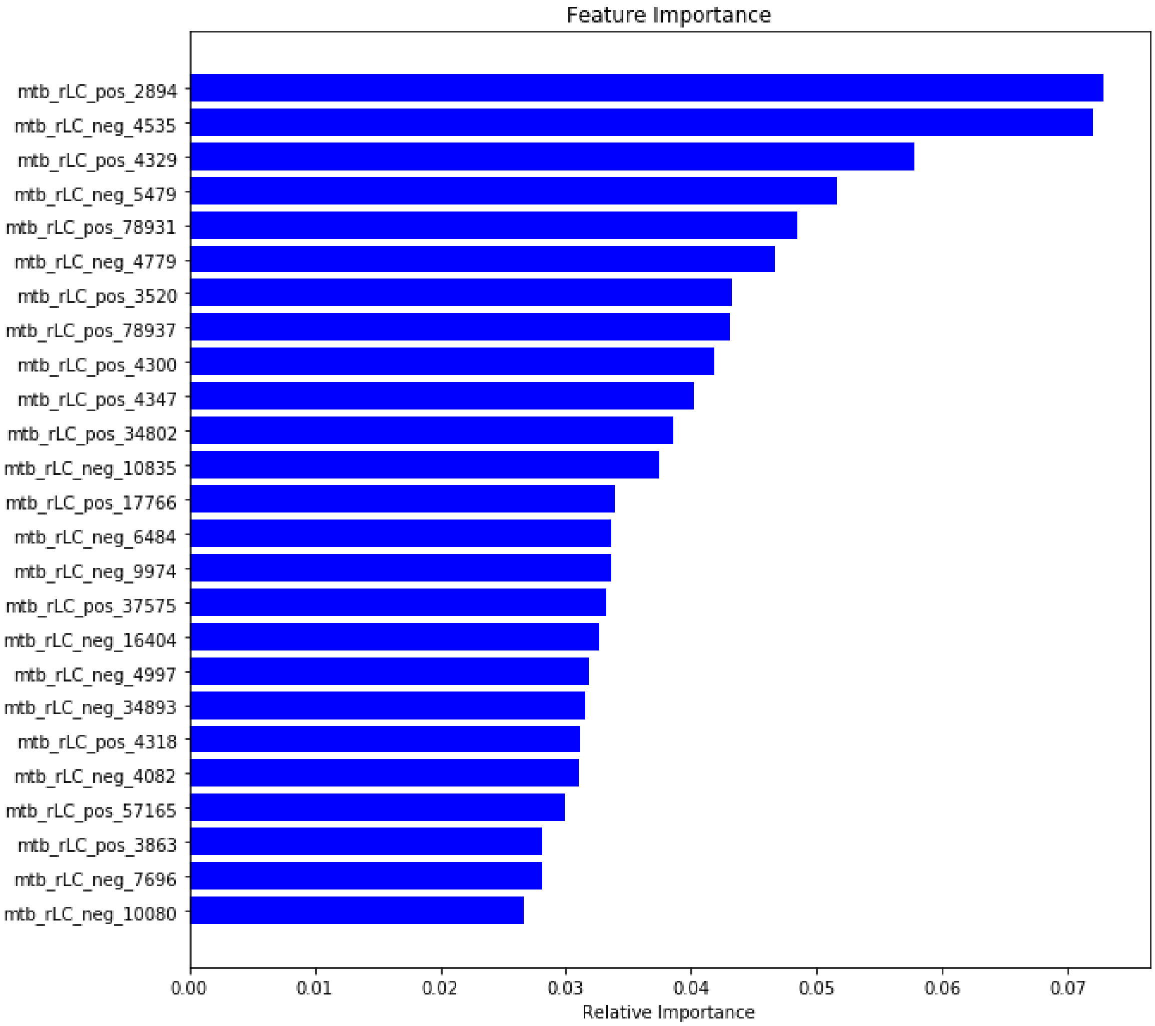

Figure 5.8 Feature importance from random forest of top 25 selected metabolites for prediction of CHD in 10 years

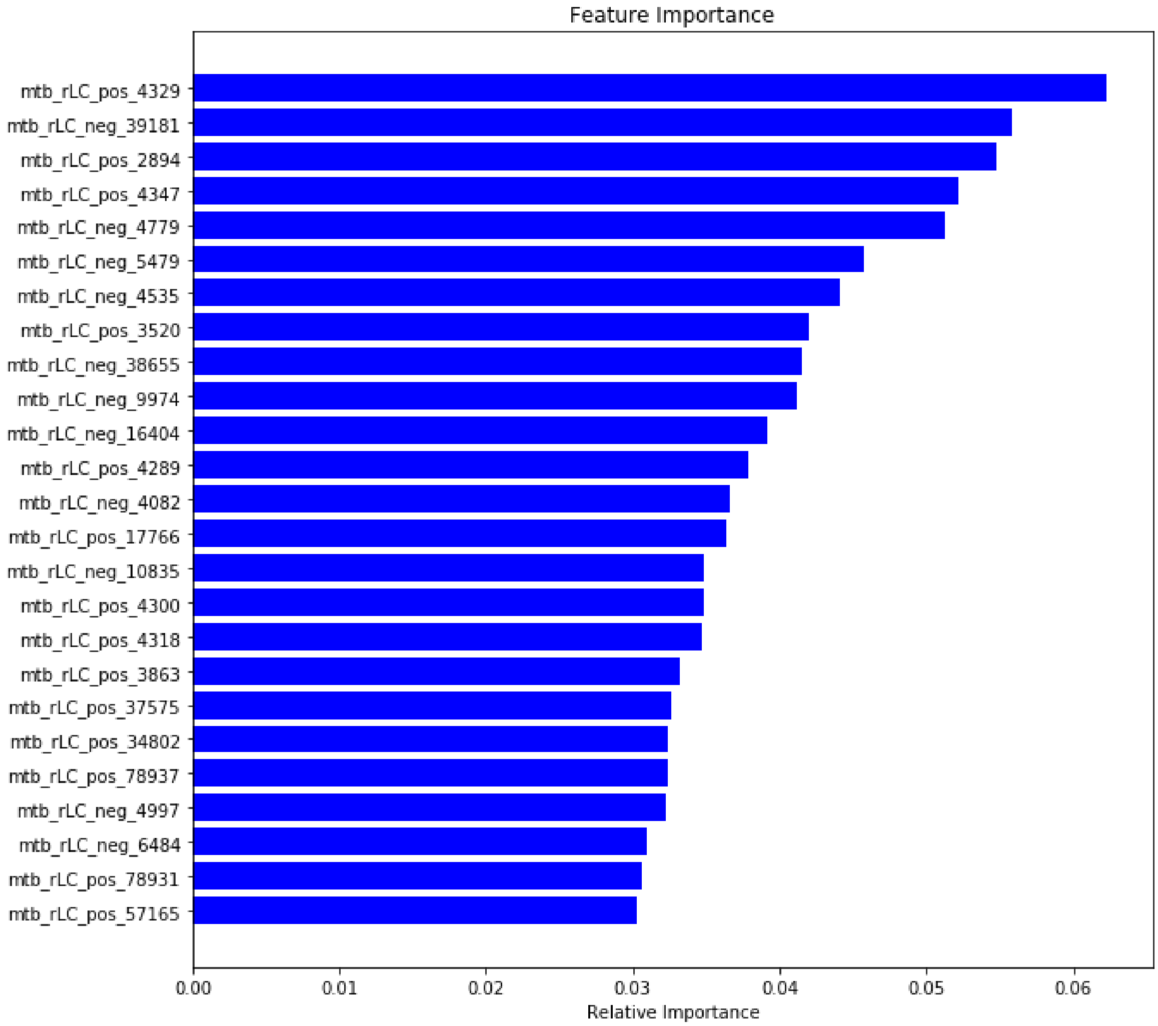

Figure 5.9 Feature importance from random forest of top 25 selected metabolites for prediction of CHD in 15 years

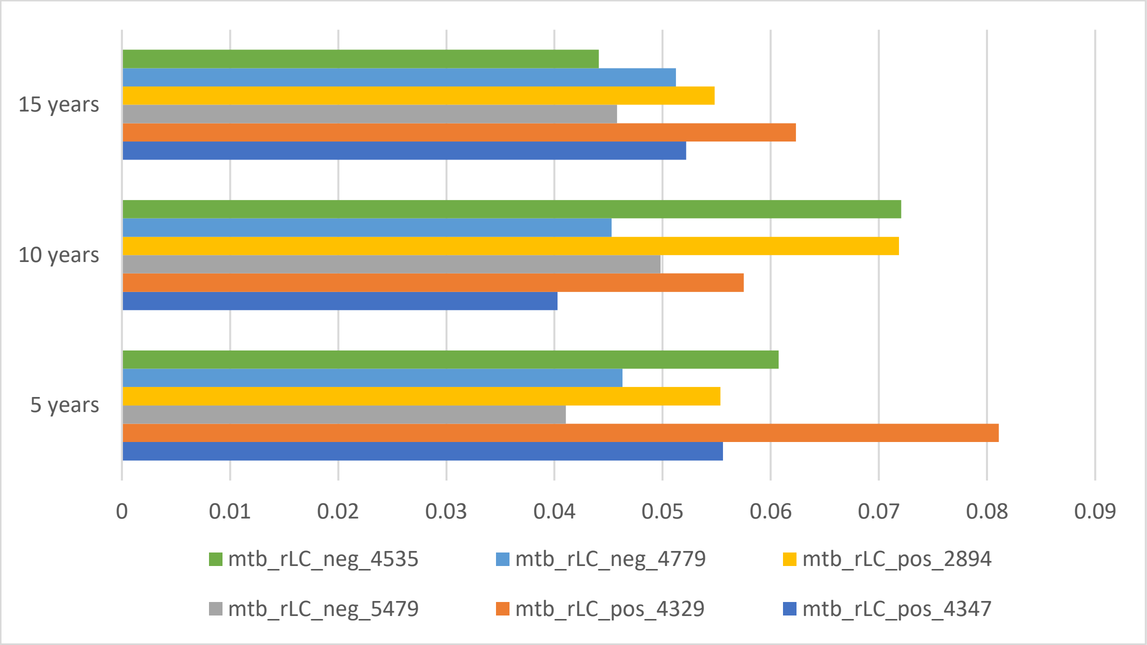

Figure 5.10 Feature importance of the common metabolites among the top 10 metabolites in each of 5, 10, 15 years

6. Discussion & Future Work

To the best of our knowledge, this is the first research of using machine learning and metabolites to predict CHD in a timely manner, which predicts the occurrence of CHD in different time frames and compares the prediction performance for those time frames. This method of prediction and comparison can provide more insights into the association between metabolites and the time progress of development of CHD. This research is an early-stage study of using metabolites and machine learning to predict CHD. However, it demonstrates that using a small number of metabolites (less than 20) can have closer prediction power in long-term, such as 15 years. The similar prediction performance for longer time could be a possible tool for lifetime prediction. This research result also show how metabolites can be associated with the development of CHD in a timely manner. This could provide insights into the understanding of the complexity of CHD and metabolic in terms of time of development. Also, because of metabolites being the end-product of gene expression, longer time prediction could also help with lifetime personalized health care. The prediction performance of metabolites also aligns with the work of Vaarhorst et al. (2014). That is, metabolites improve the prediction of CHD based on risk factors but is a promising tool with larger profile of metabolites.

This research uses an imbalanced class dataset containing 7643 instances, each of which has 247 metabolite features. The scarcity of sample size and the imbalanced class issue are both common in clinical settings. To solve the imbalanced class issue and to avoid the possibility of curse of dimensionality become more important in applying machine learning to the biomedical area. In the past decade, hundreds of algorithms have been proposed for imbalanced class data (Haixiang et al., 2017). Solving the issue of imbalanced class is complex procedure. Choosing the proper algorithm is critical to improve the performance of the classifiers. Feature selection is another important piece to enhance the model performance. An ideal set of features are relevant to the target but irrelevant with each other (Fonti & Belitser, 2017). The nature of interaction between metabolites creates challenges to select the ‘ideal’ set of features that contribute the most to classification but also have a good interpretation. The model performance could be possibly significantly improved with an integration of a more sophisticated algorithm to balance classes and an algorithm that could select harmonic features with the balanced classes.

In this research, the class label is based on the integration of the first occurrence of cardiovascular incident and the follow-up time to that incident. If more data of the physical development of CHD are accessible, such as stenosis of coronary arteries, the result could be different. Since the stenosis degree could signalize the development of CHD, metabolites may have underlying relationship with those changes. This dataset is derived from FINRISK, in which the participants are mainly white people. Some studies have shown that the current risk assessment system by using risk factors have visible errors in South Asians living in the U.K because the existing risk assessment systems are mainly based on white race (Bhopal et al., 2005). The effect of different races could be considered as a factor in the performance in using metabolites to predict as well. It can be intriguing comparison with the result of this research.

With the advancement of NRM, over 250,000 samples of metabolites could be generated annually (Soininen et al., 2015). With this fast and low-cost high throughput metabolite detection platform, numerous possibilities could be studied for and achieved by understanding the association between metabolites with disease. Low-cost in the detection of metabolites could increase the accessibility to CHD detection, which could enable earlier preventative medical intervention. Based on this research, we could see that one single measure of metabolites have the potential to have the same prediction power as risk factors. One single measure could be achieved through, for example, finger prick blood sampling. There is already commercialized home-kit finger prick sampling for testing hormones

(Everlywell, n.d). It is possible that CHD could be tested in this convenient way in the future. Metabolites also enable quantitative molecular data, which could become a promising tool for personalized health care. Combined with the power of machine learning and artificial intelligence (AI), the detection of any signal for the development of CHD could be accessible in a daily life setting. The intersection of using machine learning and metabolites to predict CHD opens countless possibilities to study the underlying molecular mechanism. It would become as a promising tool with more machine learning algorithms specialized for data like clinical data, imbalanced class and high-dimensional.

Bibliography

World Health Organization. (2014). *Global status report on noncommunicable diseases 2014* (No. WHO/NMH/NVI/15.1). World Health Organization.

Sanchis-Gomar, F., Perez-Quilis, C., Leischik, R., & Lucia, A. (2016). Epidemiology of coronary heart disease and acute coronary syndrome. *Annals of translational medicine*, *4*(13).

American Heart Association. (2013). Coronary artery disease-coronary heart disease. *Obtenida el*, *13*.

Okrainec, K., Banerjee, D. K., & Eisenberg, M. J. (2004). Coronary artery disease in the developing world. *American heart journal*, *148*(1), 7-15.

Hajar, R. (2017). Risk factors for coronary artery disease: historical perspectives. *Heart views: the official journal of the Gulf Heart Association*, *18*(3), 109.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, *97*(18), 1837-1847.

Lloyd-Jones, D. M. (2010). Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation*, *121*(15), 1768-1777.

Borodulin, K., Tolonen, H., Jousilahti, P., Jula, A., Juolevi, A., Koskinen, S., Kuulasmaa, K., Laatikainen, T., Männistö, S., Peltonen, M., Perola, M., Puska, P., Salomaa, V., Sundvall, J., Virtanen, S. M., & Vartiainen, E. (2017). Cohort Profile: The National FINRISK Study. *International Journal of Epidemiology*, *47*(3), 696–696i. https://doi.org/10.1093/ije/dyx239

Delles, C., Rankin, N. J., Boachie, C., McConnachie, A., Ford, I., Kangas, A., Soininen, P., Trompet, S., Mooijaart, S. P., Jukema, J. W., Zannad, F., Ala-Korpela, M., Salomaa, V., Havulinna, A. S., Welsh, P., Würtz, P., & Sattar, N. (2017). Nuclear magnetic resonance-based metabolomics identifies phenylalanine as a novel predictor of incident heart failure hospitalisation: results from PROSPER and FINRISK 1997. *European Journal of Heart Failure*, *20*(4), 663–673. https://doi.org/10.1002/ejhf.1076

Wang, Z., Zhu, C., Nambi, V., Morrison, A. C., Folsom, A. R., Ballantyne, C. M., Boerwinkle, E., & Yu, B. (2019). Metabolomic Pattern Predicts Incident Coronary Heart Disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *39*(7), 1475–1482. https://doi.org/10.1161/atvbaha.118.312236

Miggiels, P., Wouters, B., van Westen, G. J., Dubbelman, A. C., & Hankemeier, T. (2019). Novel technologies for metabolomics: More for less. *TrAC Trends in Analytical Chemistry*, *120*, 115323.

Vaarhorst, A. A., Verhoeven, A., Weller, C. M., Böhringer, S., Göraler, S., Meissner, A., Deelder, A. M., Henneman, P., Gorgels, A. P., van den Brandt, P. A., Schouten, L. J., van Greevenbroek, M. M., Merry, A. H., Verschuren, W. M., van den Maagdenberg, A. M., van Dijk, K. W., Isaacs, A., Boomsma, D., Oostra, B. A., van Duijn, C.M., Jukema, J.W., Boer, J.M.A., Feskens, E., Heijmans, B.T., Slagboom, P. E. (2014). A metabolomic profile is associated with the risk of incident coronary heart disease. *American Heart Journal*, *168*(1), 45–52.e7. https://doi.org/10.1016/j.ahj.2014.01.019

Marcinkiewicz-Siemion, M., Kaminski, M., Ciborowski, M., Ptaszynska-Kopczynska, K., Szpakowicz, A., Lisowska, A., Jasiewicz, M., Tarasiuk, E., Kretowski, A., Sobkowicz, B., & Kaminski, K. A. (2020). Machine-learning facilitates selection of a novel diagnostic panel of metabolites for the detection of heart failure. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-019-56889-8

Gutiérrez-Esparza, G. O., Infante Vázquez, O., Vallejo, M., & Hernández-Torruco, J. (2020). Prediction of metabolic syndrome in a Mexican population applying machine learning algorithms. *Symmetry*, *12*(4), 581

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-106

Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*, *30*, 1-25.

Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In *2017 IEEE symposium on computers and communications (ISCC)* (pp. 204-207). IEEE.

Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems (IJAIS)*, *3*(7), 25-30.

Forssen, H., Patel, R., Fitzpatrick, N., Hingorani, A., Timmis, A., Hemingway, H., & Denaxas, S. (2017). Evaluation of machine learning methods to predict coronary artery disease using metabolomic data. In *Stud Health Technol Inform* (Vol. 235, pp. 111-115). IOS Press.

Krishnani, D., Kumari, A., Dewangan, A., Singh, A., & Naik, N. S. (2019, October). Prediction of coronary heart disease using supervised machine learning algorithms. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* (pp. 367-372). IEEE.

Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.

Campbell, C., & Ying, Y. (2011). Learning with support vector machines. *Synthesis lectures on artificial intelligence and machine learning*, *5*(1), 1-95.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220-239.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, *77*(1), 103-123.

Marma, A. K., & Lloyd-Jones, D. M. (2009). Systematic examination of the updated Framingham heart study general cardiovascular risk profile. *Circulation*, *120*(5), 384.

Cavanaugh-Hussey, M. W., Berry, J. D., & Lloyd-Jones, D. M. (2008). Who exceeds ATP-III risk thresholds? Systematic examination of the effect of varying age and risk factor levels in the ATP-III risk assessment tool. *Preventive medicine*, *47*(6), 619-623.

Cuperlovic-Culf, M. (2018). Machine learning methods for analysis of metabolic data and metabolic pathway modeling. *Metabolites*, *8*(1), 4.

Bhopal, R., Fischbacher, C., Vartiainen, E., Unwin, N., White, M., & Alberti, G. (2005). Predicted and observed cardiovascular disease in South Asians: application of FINRISK, Framingham and SCORE models to Newcastle Heart Project data. *Journal of public health*, *27*(1), 93-100.

Soininen, P., Kangas, A. J., Würtz, P., Suna, T., & Ala-Korpela, M. (2015). Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circulation: cardiovascular genetics*, *8*(1), 192-206.

Everlywell. (n.d.). *Home Health Testing | Results You Can Understand*. Retrieved June 2, 2021, from https://www.everlywell.com/

Yu, C. S., Lin, Y. J., Lin, C. H., Wang, S. T., Lin, S. Y., Lin, S. H., Wu, J. L., & Chang, S. S. (2020). Predicting Metabolic Syndrome With Machine Learning Models Using a Decision Tree Algorithm: Retrospective Cohort Study. *JMIR Medical Informatics*, *8*(3), e17110. https://doi.org/10.2196/17110